



# Représentation et Stockage des données de la numérisation du dictionnaire Trévoux

Ingrid Falk

## ► To cite this version:

Ingrid Falk. Représentation et Stockage des données de la numérisation du dictionnaire Trévoux.  
[Rapport de recherche] Loria. 2006. hal-01075496

**HAL Id: hal-01075496**

**<https://inria.hal.science/hal-01075496>**

Submitted on 17 Oct 2014

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Représentation et Stockage des données de la numérisation du dictionnaire Trévoux

Ingrid Falk

15 mars 2006

## Résumé

Le dictionnaire dit de « Trévoux » est un « dictionnaire universel françois et latin » dont la première édition a été publiée en 1704. Sa numérisation en vue à la fois de la préservation et de la valorisation fait partie du programme du CPER ILD-ISTC. Dans le cadre de mes travaux pour READ j'ai exploré les outils et ressources de normalisations mise en œuvre dans d'autres projets de création de collections numériques, ce qui m'amène à proposer un mode de stockage dans des formats XML standardisés en préservant les liens entre les images scannées, les objets reconnus par les OCR et des différentes versions du contenu textuel.

## 1 Introduction

Ce rapport présente mon travail au sein de l'équipe READ dans le cadre du projet CPER ILD-ISTC. Il s'est déroulé de janvier à mars 2006. Au cours de ce travail j'ai exploré comment représenter les données provenant de la numérisation du dictionnaire Trévoux dans un souci à la fois de préservation et de souplesse d'utilisation et de présentation. Dans ce rapport je présente d'abord (très brièvement) les standards existants et utilisés dans le domaine des collections numériques. Ensuite, on montre à travers d'un exemple - l'image scannée des pages 105 et 106 du dictionnaire Trévoux - comment on pourrait procéder pour représenter et stocker les ressources en résultant dans les formats mentionnés précédemment, avec le moins d'intervention humaine possible. Finalement, je présente un exemple d'application : à partir des fichiers XML composant l'objet numérique on génère une page web (en XHTML + CSS) pour visualiser les correspondances entre des parties de l'image initiale, les objets identifiés par l'OCR et les textes codés dans le document de vérité.

## 2 Les différents standards

### 2.1 La TEI

La **TEI** (« Text Encoding Initiative »), v. [1], fondée à l'origine sur le SGML et s'appuyant désormais sur le XML, est un langage de marquage qui permet d'échanger des données textuelles, notamment pour les sciences humaines et les études sur les textes littéraires.

Mais en fait, la TEI n'est pas une DTD, mais un ensemble de recommandations (« Guidelines ») et d'éléments rassemblés en modules distincts (« tag sets ») dont l'utilisation et le choix forment une DTD particulière. Il n'existe donc pas une DTD TEI mais autant que les combinaisons de modules rendent possibles. Pour reprendre les mots de Lou Burnard (cf. [2]), la TEI est un système extensible, modulaire et polymorphe constituant un modèle abstrait.

Par exemple, la TEI convient parfaitement à la fois pour coder un document de vérité et les entrées d'un dictionnaire mais, on se servira de schéma différents :

Pour les documents de vérité on mettra l'accent plutôt sur les particularités typographiques :

- les sauts de lignes
- les marques de pages
- changements de polices
- glyphes inhabituels
- taches
- fragments illisibles

Par contre, pour le texte finalisé on voudra marquer :

- des entrées du dictionnaire
- des mots vedettes
- la description morphologique et syntaxique
- homonymes et synonymes

Les deux documents auront un set de balises de bases communes auxquelles on va rajouter celles qui conviendront le mieux au document et à l'utilisation envisagée de l'encodage.

La TEI est largement utilisée et documentée. Elle propose des formats de codage pour une large variété de textes et d'usages, mais si elle s'avère incomplète elle peut toujours être complétée.

Si la TEI a été conçu pour représenter des structures textuelles, elle ne se prête pas aussi bien à la représentation détaillée et exhaustive de la mise en page physique d'un document imprimé. Mais pour la reconnaissance et structuration automatique de documents on a besoin de toutes ces informations physiques parce qu'on voudrait en inférer des unités logiques. Par exemple, dans le dictionnaire « Trévoux » les mots vedette sont toujours typographiés en majuscule, les débuts de paragraphe sont indentés. Une fois que l'on dispose de cette information il conviendrait de la préserver, ainsi que le lien avec les images des originaux.

## 2.2 ALTO

Si la TEI est établie comme un standard pour la représentation structurée de texte, il m'a été beaucoup plus difficile de trouver un format (standardisé) pour représenter d'une manière exhaustive la composition d'une page d'un texte imprimé : en marges et un espace imprimé et en blocs physiques - textuels, illustrations ou composés.

**ALTO** (Analysed layout and text object), cf. [3], est le seul format ouvert et assez générique que j'aie trouvé pour représenter les résultats des OCR sur du texte imprimé.

C'est un schéma XML issu du projet européen **METAE** (Metadata Engine Project - cf. [7]) qui est actuellement utilisé aussi par le **National Digital Newspaper Program** aux Etats Unis (cf. [8]).

## 2.3 Putting it all together : le format de métadonnées METS

**METS** (Metadata Encoding and Transmission Standard), v. [4], est un schéma XML élaboré à l'initiative de la Digital Library Federation. Il sert à encoder les métadonnées descriptives, administratives et structurelles de documents, ou plutôt d'objets numériques. Dans cette vision un objet numérique sera composé de plusieurs documents numériques (images, sons, textes en différents formats ...) et un fichier METS répertoriant les noms et la localisation des fichiers leur correspondant, des métadonnées descriptives et administratives, mais aussi une carte de structure pouvant représenter des correspondances entre des sousensembles de ces documents.

METS est déjà utilisé dans beaucoup de projets de numérisation, entre autre par l'École des chartes (cf. [9]) et la Bibliothèque Nationale de France.

## 3 Application au dictionnaire Trévoux

**Les objets graphiques** Dans cet exemple l'objet numérique sera construit autour de l'image scannée des pages 105 et 106 du dictionnaire Trévoux (édition ??). Le document numérique initial est un fichier image (DUT01\_0082.tif) dans le format tif : (1).

En vue d'une présentation sur internet on pourrait convertir cette image en d'autres formats (comme png, gif ou jpeg) ou encore de prévoir des images plus petites en taille.

**L'analyse physique** L'OCR va analyser l'image et va stocker ses résultats dans un format XML propriétaire. Ce fichier XML pourra être converti (automatiquement) dans un fichier XML alto DUT01\_0082\_alto.xml (2)

**Les formats textuels** Pour pouvoir analyser les résultats de l'OCR on a besoin d'un document de vérité. Ce document doit être produit manuellement, mais il peut être codé dans un format TEI approprié : DUT01\_0082\_tei\_3.xml (3).

**Les métadonnées pour monter l'ensemble** Finalement on va regrouper dans un fichier METS les informations sur

- les fichiers composants cet objet numérique (4) :
  - les images : DUT01\_0082.tif, DUT01\_0082.jpg ...
  - le fichier ALTO avec l'analyse (physique) de l'OCR : DUT01\_0082\_alto.xml
  - le document de vérité (en TEI) : DUT01\_0082\_tei\_3.xml
- des liens structurels entre ces documents et quelquesunes de leurs sous-divisions (5)

L'illustration 6 montre schématiquement le type d'informations que peut représenter un fichier METS et le fonctionnement de la carte structurale.

Le fichier METS peut être produit automatiquement (par des scripts ou transformation XSLT).

**L'objet numérique** Pour conclure, dans ce cas, l'objet numérique sera constitué :

- des fichiers images : DUT01\_0082.tif, DUT01\_0082.jpg, ... (cf. fig 1)
- du document de vérité (s'il y en a) : DUT01\_0082\_tei\_3.xml (cf. fig 3)

blir un à Constantinople, qui après la mort du Saint fut transféré en Bithynie par Jean son successeur. A Jean succéda Marcellus, que Nicephore a cru être l'Instituteur des *Acoemètes*. Sous ce Marcellus ce pieux institut s'étendit beaucoup, dit Bollandus; & c'est là apparemment ce qui a fait que Nicéphore l'en a cru fondateur. Ce fut de son tems que Studius vint de Rome à Constantinople y bâtir un Monastère, & y mit des Moines, qu'il tira des Monastères *Acoemètes*. Ce fut là l'origine des Studites, qui conséquemment viennent des *Acoemètes*. Saint Jean Calybite se retira dans un Monastère d'*Acoemètes*; & non pas d'*Aromètes*, comme a imprimé la Saüllye dans le Martyrologe de France. Quoique les *Acoemètes* aient fleuri sur tout en Orient, il y en a cependant eu quelques-uns en Occident. Le P. le Coindre prétend à l'endroit que je cite, qu'il n'y a eu que le Monastère de Luxeuil, *Luxovienſe*, celui de Remiremont, *Abendense*, & celui de S. Salaberge à Laon, où l'on ait dit perpétuellement l'Office de la manière que nous l'avons expliqué. Le P. Mabillon soutient qu'il y faut ajouter celui de S. Maurice, *agavinſe*, fondé par Sigismond Roi de Bourgogne, celui de S. Marcel de Châlons, & celui de Saint Denys en France. D'autres ajoutent encore celui de S. Riquier &c. Il n'est pas vrai que S. Eucher Evêque d'Orléans se fit Moine *Acoemète*, comme l'a dit Canisius. Ce fut dans un Monastère de Bénédictins, à J. lieux de Rouen, qu'il se retira, comme l'a remarqué Bollandus. T. 1. de Janv. p. 1019.

On a aussi appelé *Acoemètes* les Stylites, & quelques autres Moines de la Palestine, mais dont l'institut étoit fort différent de celui des *Acoemètes*. On pourroit aujourd'hui appeler *Acoemètes* les Religieuses du S. Sacrement, qui ont l'adoration perpétuelle, & se relèvent jour & nuit, en sorte qu'il y en ait toujours devant le S. Sacrement à prier.

Entre Nicéphore & Bollandus, dont j'ai parlé, Théodore Lècteur, I. 1. Evagrius I. 121<sup>re</sup>, C. 18. & 21. Théopane, Cédrenus, l'Auteur de la vie de S. Alexandre dans Bolland. 15. Janv. & Jacobus Canisius dans le Ribadensis Latin au 20. Février, Baronius à l'an 459. M. du Fresne dans son Glossaire, le Coindre Annal. T. 1. an. 536. n. 224, & suiv. Le P. Mabillon *Ad. Sanct. Bened. Sac. IV. p. 2. Pref.* ont écrit des *Acoemètes*.

**ACOINT**, *tr. s. adj.* Ce mot veut dire familial, selon Nicod. *Amicus familiaris*.

**ACOLAIAN**, *f. m.* Insecte de l'Isle de Madagascar. Il ressemble à une punaise. Il est plus gros.

**ACOLYTHAT**, *f. m.* *Acolythus*. Ordre, rang d'*Acolythe*: c'est le premier des moindres Ordres; c'est-à-dire; celui qui précède immédiatement le Soudiaconat.

**ACOLYTHE**, *f. m.* Terme Ecclésiastique. *Acolythus*. Les Grecs donnoient ce nom à ceux qui étoient inébranlables dans leurs résolutions. C'est par cette raison que les Stoïciens furent appelés *Acolythes*; parce que rien ne pouvoit leur arracher leurs sentimens. Ils trouvoient même qu'il y avoit de la lâcheté à en changer. Depuis, l'Eglise Chrétienne a consacré ce nom, en l'appliquant à ceux qui se devoient au service de Dieu. Anciennement les jeunes gens qui alloient au ministère Ecclésiastique, accompagnoient & suivoient les Evêques par tout, soit pour les servir, soit pour être les témoins de leur conduite. Cette assiduité à suivre les Evêques les fit appeler *Acolythes*. Saint Cyprien dit lui-même, qu'il avoit des *Acolythes*. Aujourd'hui les fonctions des *Acolythes* sont bien différentes de la première institution. Un *Acolythe* est celui qui a seulement reçu le premier & le plus considérable des quatre Ordres Mineurs dans l'Eglise; dont l'emploi est d'allumer les cierges, de porter les chandeliers, la navette où est l'encens, de préparer le vin & l'eau pour le sacrifice, & de rendre d'autres services à l'autel. Autrefois les *Acolythes* ramassoient dans un sac ce que les fidèles avoient offert, & qui avoit été béni pendant la messe, & après qu'elle étoit finie ils le donnoient aux Prêtres, qui devoient le diviser. Le devoir des *Acolythes* est d'accompagner l'Evêque, ou le Prêtre, & de leur rendre service dans les fonctions Ecclésiastiques. C'est l'Ordre que les jeunes Clercs exercent le plus. Il y avoit à Rome trois sortes d'*Acolythes*. Les *Acolythes* du Palais, *Palatini*, qui servoient le Pape; les *Acolythes* Stationnaires, *Stationarii*, qui servoient dans les Eglises, où il y avoit Station, les *Acolythes* Régionnaires, *Regionarii*, qui servoient avec les Diacres dans les différens quartiers de la ville. On trouve aussi des *Acolythes* parmi les Officiers Auliques de Constantinople; & Curopalates dit que le Capitaine, ou Chef de la Cohorte Impériale de Bizance, étoit nommé *Acolythe*.

Dans l'Euchologe des Grecs on trouve les leçons qu'on lit lorsqu'on ordonne des Lècteurs; mais il n'y est point parlé des autres moindres Ordres, qui sont, l'Ordre de Portier, d'Exorciste, & d'*Acolythes*: ce qui pourroit faire croire que les Grecs ne confèrent point ces Ordres-là aujourd'hui. Le Père Goar

dans ses notes sur l'Euchologe, répond qu'on ne peut pas douter que ces trois moindres Ordres n'aient été connus de l'ancienne Eglise Grecque, & qu'elle n'ait eu des Ministres qui les avoient, puisqu'il en est fait mention dans Saint Denis, S. Ignace Martyr, S. Saint Epiphane, dans les Conciles de Laodicée & d'Antioche, dans les Nouvelles de Juslinien, dans Phatius &c. Il ajoute, qu'il semble que les Grecs d'aujourd'hui ont des *Acolythes* sous le nom de Députés & de Cérofétaire. Les Missionnaires Latins qui sont en Grèce disent que les Grecs ont aujourd'hui des *Acolythes*, & les autres moindres Ordres; & on doit plus les en croire que le Père Martène, qui assure que l'Ordre des *Acolythes* a été tout à fait inconnu à l'Eglise d'Orient. Voyez le P. Goar sur l'Euchologe; le P. Martène sur le Pontifical, l'Ordre Romain, &c.

Ce mot vient du Grec *ἀκόλῳτος*, qui signifie suivre. Et *Acolythe*, un suivant. C'est ainsi que l'explique le Glossaire Grec & Latin; & Macêr; mais Dominique son frère le tire de l'a privatif, & de *κόλῳ*, empêcher.

**ACOMAS**, *f. m.* Arbre qui croît dans les Isles Antilles; & dont le bois s'emploie aux ouvrages de Menuiserie. Cet arbre est à peu près de la hauteur de nos pommiers; ses feuilles sont assez longues & lisses; son fruit est de la grosseur d'une prune, il devient jaune dans la maturité. Son amertume empêche qu'on ne le mange; il n'y a que les pigeons ramiers qui puissent s'accommoder de son fruit; mais leur chair en retient si fort le goût, qu'on ne peut les manger dans le tems qu'ils s'en nourrissent; l'écorce de cet arbre est raboteuse, cendrée, & elle donne un suc laiteux lorsqu'on l'incise. Son bois est pesant, de couleur rouge, tirant sur le jaunâtre; le cœur est d'un rouge tirant sur le violet. Ces couleurs varient suivant son âge; & tout le bois prend fort bien le poli. ROCHERFORT. Le Père du Tertre rapporte qu'un Nègre l'avoit guéri d'un grand mal de dents en lui frottant les tempes & le derrière des oreilles avec le lait qui se tire de l'écorce de l'*Acomas* franc. Car ce Père; Histoire des Antilles, Traité 3. C. 4. §. 3. distingue trois sortes d'*Acomas*; l'*Acomas* franc, qui est un des plus gros; & des plus hauts arbres des Antilles, & le meilleur de tous pour les bâtimens; l'*Acomas* bâtard, qui croît à la Capsterre de la Guadeloupe, qui n'est ni si beau, ni si bon à bâtir que le précédent; & le troisième, qui croît aux environs de la grande Ance, semblable au premier, sinon que le cœur en est rouge.

**ACOMMICHER**, *Verb. act.* Vieux mot François, qui vouloit dire *Communier*, donner la Communion. En fit le Roi d'Orléans grand planté de Mèsses, pour *accommoder* ceux qui devoient en avoir. FROISSARD.

**ACOMPARAGER**, *Verbe act.* Ce mot, selon Nicod, veut dire *comparer*. *Conferre*, *comparare*.

**ACOMSICT**, *part. & adj.* Ce mot dans l'Ercéval veut dire *pour suivre*.

**ACON**, *f. m.* *Cymba*. Terme de marine. Petit bateau à fond plat, dont on se sert pour aller sur la vase quand la mer est retirée.

**ACONIT**, *f. m.* *Aconitum*. n. Plante venimeuse. Les anciens Botanistes ont attribué ce nom à plusieurs plantes de différens genres. Celles dont il s'agit ici sont leurs fleurs irrégulières, composées de plusieurs pétales, dont l'assemblage représente assez bien un casque ouvert; c'est-à-dire, que la pétale supérieure fait le casque du heaume, les deux latérales tiennent la place des deux oreillettes, & les inférieures représentent la mentonnière. Les espèces qu'on nomme tué-loup, *Lycostomum*, *Λυκοστόμον*, ont leur casque allongé en manière de toque, ou de bonnet à la Polonoise. Les fruits qui succèdent aux fleurs sont composés de plusieurs graines, qui s'ouvrent selon leur longueur, & renferment des semences anguleuses, & chagrinées. Ses feuilles sont arrondies & découpées plus ou moins profondément. Ce genre d'*aconit* comprend plusieurs espèces; qu'on peut ranger sous trois principales classes. La première est de celle dont toute la fleur est bleue, ou violette, & la pétale supérieure de la fleur forme un casque. On la nomme Napel, *Napellus*, à *Napo*, à cause que ses racines sont en naves. Le Napel est très-dangereux; mais on a trop exagéré son venin. La seconde est de celle qui a ses fleurs tout à fait semblables à celles du Napel; hormis qu'elles sont jaunes. Elle s'appelle Anthora, *Anti-Thora*. C'est-à-dire, plante souveraine contre les mauvais effets du Thora. Elle est aussi venimeuse que le Napel. Il est faux que l'Anthora croisse toujours auprès du Thora, ou du Napel. L'*aconit* de la troisième classe se distingue des deux précédentes par la figure allongée de son casque. Ses fleurs sont pâles, ou jaunâtres. On l'a appelée tué-loup, étrangle-loup, tué-chien, à cause de ses effets. *Aconitum Lycostomum*, *Λυκοστόμον*, *Κονοστόμον*. La première & la dernière de ces trois sortes d'*aconit* sont très-caustiques, très-âcres, & causent des convulsions mortelles, ou des inflammations suivies d'une gangrène prochaine.

FIG. 1 – image scannée des pages 105 et 106 du dictionnaire Trévoux

```

<?xml version="1.0" encoding="UTF-8"?>
<alto>
  <Description/>
  <Styles>
    <TextStyle ID="Arial_11." FONTFAMILY="Arial" FONTSIZE="11."/>
    <TextStyle ID="Times New Roman_10._466_29" FONTFAMILY="Times New Roman" FONTSIZE="10." FONTSPACING="29" FONTSCA-
LING="466"/>
  </Styles>
  <Layout>
    <Page WIDTH="2480" HEIGHT="2480" ID="P0_DUT01_0082" PHYSICAL_IMG_NR="1">
      <PrintSpace>
        <TextBlock ID="B0" HPOS="64" WIDTH="124" VPOS="54" HEIGHT="74">
          <TextLine BASELINE="124" ID="B0_L0" HPOS="72" WIDTH="101" VPOS="58" HEIGHT="66" STYLEREFS="Times New
Roman_22._616_-17">
            <String CONTENT="'%5" STYLE="bold italic" HPOS="72" VPOS="58" WIDTH="101"/>
          </TextLine>
        </TextBlock>
        <TextBlock ID="B3" HPOS="52" WIDTH="1174" VPOS="136" HEIGHT="4038">
          <TextLine BASELINE="189" ID="B3_L0" HPOS="98" WIDTH="1115" VPOS="144" HEIGHT="58" STYLEREFS="Times New
Roman_11.__-4">
            <String CONTENT="■blir" HPOS="98" VPOS="186" WIDTH="80"/>
            <SP HPOS="178" VPOS="161" WIDTH="10"/>
            <String CONTENT="un" HPOS="188" VPOS="161" WIDTH="43"/>
            <SP HPOS="231" VPOS="148" WIDTH="10"/>
            <String CONTENT="d" HPOS="241" VPOS="148" WIDTH="17"/>
            <SP HPOS="258" VPOS="148" WIDTH="10"/>
            <String CONTENT="Cqntantinppl" HPOS="268" VPOS="148" WIDTH="273"/>
            <String CONTENT="'txanf" HPOS="1111" VPOS="156" WIDTH="89" SUBS.TYPE="HypPart1">
              <HYP CONTENT="-" HPOS="1200" VPOS="174" WIDTH="13"/>
            </String>
          </TextLine>
          <TextLine BASELINE="237" ID="B3_L1" HPOS="89" WIDTH="1123" VPOS="192" HEIGHT="61" STYLEREFS="Times New
Roman_11.__1">
            </TextLine>
          </TextBlock>
        </PrintSpace>
      </Page>
    </Layout>
  </alto>

```

FIG. 2 – extrait du fichier ALTO

```

<?xml version="1.0" encoding="iso-8859-1"?>
<!DOCTYPE TEI PUBLIC "-//TEI P5//DTD Main Document Type//EN" "Transcription-
dictionnaires-anciens.dtd">
<TEI xmlns="http://www.tei-c.org/TEI_P5/">
  <teiHeader>
    <fileDesc><!-- ..... --></fileDesc>
    <encodingDesc>
      <tagsDecl>
        <rendition xml:id="sup">text is placed higher (superscript)</rendition>
        <tagUsage gi="hi">used to mark typographic specificities</tagUsage>
      </tagsDecl>
      <charDesc>
        <!-- description of specific typographic glyphs -->
        <glyph xml:id="longS">
          <glyphName>LATIN SMALL LETTER LONG S</glyphName>
          <mapping type="PUA">U+017f</mapping>
          <desc>in common use in Roman types until the 18th century,
used here for the sound 's'</desc>
          <graphic url="xx"/>
        </glyph>
        <!-- .... -->
      </charDesc>
    </encodingDesc>
  </teiHeader>
  <text>
    <body>
      <div>
        <pb n="image_DUT01_0082" xml:id="image_DUT01_0082_p0"/>
        <fw type="head" place="top-centre">ACO.</fw><lb xml:id="l0"/>
        <fw type="pageno" place="top-left">105</fw><lb xml:id="l1"/>
        <p xml:id="par0">
          blir un à Con<g ref="#longST">st</g>antinople, qui après la mort du
          Saint fut tran-<lb xml:id="l2"/>féré en Bithynie par Jean
          <g ref="#longS">s</g>on <g ref="#longS">s</g>ucce<g ref="#ss">ss</g>eur.
          A Jean <g ref="#longS">s</g>uccéda Mar-<lb xml:id="l3"/>cellus,
          que Nicephore a cru être l'In<g ref="#longST">st</g>ituteur des
          <hi rend="#italic">Acoemètes</hi>.Sous<lb xml:id="l4"/>
          <!-- ..... -->
        </p>
      </div>
    </body>
  </text>
</TEI>

```

FIG. 3 – extrait d'un document de vérité, codé en TEI

```

<?xml version="1.0" encoding="iso-8859-1"?>
<mets xmlns:xm1ns="http://www.loc.gov/METS/"
      xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
      xmlns:xlink="http://www.w3.org/1999/xlink">
  <fileSec>
    <fileGrp ID="3_ALTO" USE="alto generated automatically via OCR">
      <file ID="ALTO_DUT01_0082" MIMETYPE="text/xml">
        <FLocat LOCTYPE="URL" xlink:href="file:///home/falk/read/tei/DUT01_0082_alto.xml"/>
      </file>
    </fileGrp>
    <fileGrp ID="1_TEI" USE="Documents de vérité en TEI level 3">
      <file ID="TEI_3" MIMETYPE="text/xml">
        <FLocat LOCTYPE="URL" xlink:href="file:///home/falk/read/tei/DUT01_0082_tei_3.xml"/>
      </file>
    </fileGrp>
    <fileGrp ID="22_JPG" USE="Jpeg image">
      <file ID="JPG_DUT01_0082" MIMETYPE="image/jpeg">
        <FLocat LOCTYPE="URL" xlink:href="file:///home/falk/read/images_inge/DUT01_0082.jpg"/>
      </file>
    </fileGrp>
    <fileGrp ID="21_TIF" USE="Scanned image (tiff)">
      <file ID="TIF_DUT01_0082" MIMETYPE="image/tiff">
        <FLocat LOCTYPE="URL" xlink:href="file:///home/falk/read/images_inge/DUT01_0082.tif"/>
      </file>
    </fileGrp>
  </fileSec>
<!-- ..... -->
</mets>

```

FIG. 4 – extrait d'un fichier METS, représentation des composants de l'objet numérique



```

<?xml version="1.0" encoding="iso-8859-1"?>
<mets xmlns:xm1ns="http://www.loc.gov/METS/">
  <structMap TYPE="logical">
    <div ID="MAP_DUT01_0082" LABEL="Numérisation du dictionnaire Trévoux">
      <fptr FILEID="ALTO_DUT01_0082"/>
      <fptr FILEID="TIF_DUT01_0082"/>
      <fptr FILEID="JPG_DUT01_0082"/>
      <fptr FILEID="TEI_3"/>
    <div ID="MAP_DUT01_0082_P0" LABEL="pages 105-106"
      TYPE="representation of the scanned image of one (or two) pages">
      <fptr FILEID="TIF_DUT01_0082"/>
      <fptr FILEID="JPG_DUT01_0082"/>
      <fptr>
        <area FILEID="TEI_3" BETYPE="IDREF" BEGIN="image_DUT01_0082_p0"/>
      </fptr>
      <fptr>
        <area FILEID="ALTO_DUT01_0082" BETYPE="IDREF" BEGIN="P0_DUT01_0082"/>
      </fptr>
    <div ID="MAP_DUT01_0082_L0" LABEL="line 0, pages 105-106"
      TYPE="representation of line 0 as recognized by the OCR">
      <fptr>
        <area FILEID="TIF_DUT01_0082" SHAPE="RECT" COORDS="72,58,101,66"/>
        <area FILEID="JPG_DUT01_0082" SHAPE="RECT" COORDS="72,58,101,66"/>
      </fptr>
      <fptr>
        <area FILEID="TEI_3" BETYPE="IDREF" BEGIN="l0"/>
      </fptr>
      <fptr>
        <area FILEID="ALTO_DUT01_0082" BETYPE="IDREF" BEGIN="B0_L0"/>
      </fptr>
    </div>
    <fptr>
      <area FILEID="TIF_DUT01_0082" SHAPE="RECT" COORDS="518,45,320,77"/>
      <area FILEID="JPG_DUT01_0082" SHAPE="RECT" COORDS="518,45,320,77"/>
    </fptr>
    <fptr>
      <area FILEID="TEI_3" BETYPE="IDREF" BEGIN="l1"/>
    </fptr>
    <fptr>
      <area FILEID="ALTO_DUT01_0082" BETYPE="IDREF" BEGIN="B1_L0"/>
    </fptr>
  </div>
<!-- .... -->
</div>
</structMap>
</mets>

```

- d'un ou plusieurs documents alto représentant les résultats d'un ou des OCR :  
DUT01\_0082\_alto.xml (cf. fig 2)
- du documents (textuel) final représentant le dictionnaire au niveau des entrées lexicales.
- du fichier METS DUT01\_0082\_mets.xml (cf. fig 4, fig 5 et fig 6) qui va répertorier :
  - les objets physiques (fichiers) et leurs localisations
  - des liens structurels entre ces objets

**Un exemple d'utilisation** Une fois l'objet numérique mis en place on peut assez facilement (automatiquement, même à la demande) générer une présentation web tel que la suivante :

[http://www-int.loria.fr/~falk/read/DUT01\\_0082.html](http://www-int.loria.fr/~falk/read/DUT01_0082.html)

La page web a l'image scannée en fond d'image, quand on bouge la souris sur l'image on voit apparaître des cadres représentant les blocs-lignes reconnus par l'OCR. La partie supérieure de ces cadres contiendra le texte de la ligne tel qu'il a été reconnu par l'OCR, la ligne en dessous une ligne du document de vérité. L'alignement s'est fait automatiquement, il est normal qu'il y ait des décallages entre le texte OCR et le texte du document de vérité.

Ce n'est qu'une page web en XHTML+CSS, elle peut donc être visualisée par n'importe quel browser (conforme standard). Ceci n'est qu'un exemple, beaucoup d'autres applications pouvant être mises en œuvre sur ce modèle.



## Références

- [1] *The Text Encoding Initiative* Manuels, tutoriels, outils pour le codage en TEI.  
<http://www.tei-c.org>.
- [2] *Digital texts with XML and the TEI*, Text Encoding Initiative  
<http://www.tei-c.org/Talks/OUCS/2004-02/One/teixml-one.pdf>.
- [3] *Analyzed Layout and Text Object* Références et détails techniques sur le format ALTO.  
<http://www.ccs-gmbh.com/alto/>
- [4] *Metadata Encoding and Transmission Standard* Site officiel de METS, avec les spécifications, présentations et exemples d'usages.  
<http://www.loc.gov/standards/mets/>
- [5] Un article sur METS dans le Bulletin d'information francophone sur l'EAD, n°19 mars 2005.  
<http://www.archivesdefrance.culture.gouv.fr/fr/publications/dafbulead19.html#ancree4>
- [6] *Métadonnées pour les nuls, épisode 3 : METS* Un blog francophone sur METS.  
<http://blogokat.canalblog.com/archives/2005/06/20/589285.html>
- [7] *Meta Data Engine* Le site du projet européen METAe.  
<http://meta-e.aib.uni-linz.ac.at/>
- [8] *National Digital Newspaper Program* Un projet de numérisation qui utilise METS et ALTO.  
<http://www.loc.gov/ndnp/>
- [9] *Les cartulaires numérisés d'Ile de France* Un autre exemple de METS et TEI mises en œuvre par l'École de Chartes.  
<http://elec.enc.sorbonne.fr/cartulaires>